

UNCLASSIFIED

Defense Technical Information Center
Compilation Part Notice

ADP013757

TITLE: Approximation by Perceptron Networks

DISTRIBUTION: Approved for public release, distribution unlimited

This paper is part of the following report:

TITLE: Algorithms For Approximation IV. Proceedings of the 2001
International Symposium

To order the complete compilation report, use: ADA412833

The component part is provided here to allow users access to individually authored sections of proceedings, annals, symposia, etc. However, the component should be considered within the context of the overall compilation report and not as a stand-alone technical report.

The following component part numbers comprise the compilation report:

ADP013708 thru ADP013761

UNCLASSIFIED

Approximation by perceptron networks

Věra Kůrková

*Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, P.O. Box 5, 182 07 Prague 8, Czechia
vera@cs.cas.cz*

1 Introduction

The classical perceptron proposed by Rosenblatt [22] as a simplified model of a neuron computes a weighted sum of its inputs and after comparing it with a threshold, applies an activation function representing a rate of neuron firing. To model this rate, Rosenblatt used the Heaviside discontinuous threshold function, which still is, together with its various continuous approximations, the most widespread type of activation used in neurocomputing. Formally, a perceptron with the Heaviside activation function computes a characteristic function of a half-space of \mathcal{R}^d , which is for practical reasons (all inputs are bounded) restricted to a box, usually $[0, 1]^d$. Thus theoretical study of perceptron networks leads to various questions concerning approximation of functions by a special class of plane waves formed by linear combinations of characteristic functions of half-spaces (corresponding to the simplest model of perceptron network called the one-hidden-layer network with a linear output unit).

Although Rosenblatt's model was inspired biologically, plane waves (sometimes called ridge functions) have been studied for a long time by mathematicians motivated by various problems from physics. In contrast to integration theory, where functions are approximated by linear combinations of characteristic functions of boxes (simple functions), the theory of perceptron networks studies approximation of multivariable functions by linear combinations of characteristic functions of half-spaces. Expressions in terms of such functions exhibit the strength and weakness of plane waves methods described by Courant and Hilbert [4], page 676: "But always the use of plane waves fails to exhibit clearly the domains of dependence and the role of characteristics. This shortcoming, however, is compensated by the elegance of explicit results."

In this paper we survey our recent results on properties of approximation by linear combinations of characteristic functions of half-spaces. We focus on existence of best approximation, impossibility of choosing among best approximations a continuous one, estimates of rates of approximation by linear combinations of n characteristic functions of half-spaces and integral representation as a linear combination of a continuum of half-spaces.

This work was partially supported by GA ČR 201/99/0092 and 201/02/0428.

2 Preliminaries

A perceptron with an activation function $\vartheta : \mathcal{R} \rightarrow \mathcal{R}$ (where \mathcal{R} denotes the set of real numbers) computes real-valued functions on $\mathcal{R}^d \times \mathcal{R}^{d+1}$ of the form $\psi(\mathbf{v} \cdot \mathbf{x} + b)$, where $\mathbf{x} \in \mathcal{R}^d$ is an input vector, $\mathbf{v} \in \mathcal{R}^d$ is an input weight vector and $b \in \mathcal{R}$ is a bias.

The most common activation functions are sigmoidals, i.e., functions with an ess-shaped graph. Both continuous and discontinuous sigmoidals are used. Here, we study networks based on the discontinuous Heaviside function ϑ defined by $\vartheta(t) = 0$ for $t < 0$ and $\vartheta(t) = 1$ for $t \geq 0$. Let H_d denote the set of functions on $[0, 1]^d$ computable by Heaviside perceptrons, i.e.,

$$H_d = \{f : [0, 1]^d \rightarrow \mathcal{R} \mid f(\mathbf{x}) = \vartheta(\mathbf{v} \cdot \mathbf{x} + b), \mathbf{v} \in \mathcal{R}^d, b \in \mathcal{R}\}.$$

Notice that H_d is the set of characteristic functions of half-spaces of \mathcal{R}^d restricted to $[0, 1]^d$.

For all positive integers d , H_d is compact in $(\mathcal{L}_p([0, 1]^d), \|\cdot\|_p)$ with $p \in [1, \infty)$ (see, e.g., [8]). This can be verified easily once the set H_d is reparameterized by elements of the unit sphere S^d in \mathcal{R}^{d+1} . Indeed, a function $\vartheta(\mathbf{v} \cdot \mathbf{x} + b)$, with a non-zero vector $(v_1, \dots, v_d, b) \in \mathcal{R}^{d+1}$, is equal to $\vartheta(\hat{\mathbf{v}} \cdot \mathbf{x} + \hat{b})$, where $(\hat{v}_1, \dots, \hat{v}_d, \hat{b}) \in S^d$ is obtained from $(v_1, \dots, v_d, b) \in \mathcal{R}^{d+1}$ by normalization.

The simplest type of multilayer feedforward network has one hidden layer and one linear output. Such networks with Heaviside perceptrons in the hidden layer compute functions of the form

$$\sum_{i=1}^n w_i \vartheta(\mathbf{v}_i \cdot \mathbf{x} + b_i),$$

where n is the number of hidden units, $w_i \in \mathcal{R}$ are output weights and $\mathbf{v}_i \in \mathcal{R}^d$ and $b_i \in \mathcal{R}$ are input weights and biases, respectively. The set of all such functions is the set of all linear combinations of n elements of H_d and is denoted by $\text{span}_n H_d$.

For all positive integers d , $\bigcup_{n \in \mathcal{N}_+} \text{span}_n H_d$ (where \mathcal{N}_+ denotes the set of all positive integers) is dense in $(\mathcal{C}([0, 1]^d), \|\cdot\|_c)$, the linear space of all continuous functions on $[0, 1]^d$ with the supremum norm, as well as in $(\mathcal{L}_p([0, 1]^d), \|\cdot\|_p)$ with $p \in [1, \infty]$ (see, e.g., [5, 9]).

3 Existence of a best approximation

A subset M of a normed linear space $(X, \|\cdot\|)$ is called proximal if for every $f \in X$ the distance $\|f - M\| = \inf_{g \in M} \|f - g\|$ is achieved for some element of M , i.e., $\|f - M\| = \min_{g \in M} \|f - g\|$ (see, e.g., [23]). Clearly, a proximal subset must be closed.

A sufficient condition for proximality of a subset M of a normed linear space $(X, \|\cdot\|)$ is compactness or bounded compactness. However, by extending H_d into $\text{span}_n H_d$ for any positive integer n we lose compactness. Nevertheless compactness can be replaced by a weaker property that requires only those sequences that "minimize" a distance from M of an element of X to have convergent subsequences. More precisely, a subset M of a normed linear space $(X, \|\cdot\|)$ is called *approximatively compact* if for each $f \in X$ and any sequence $\{g_i : i \in \mathcal{N}_+\} \subseteq M$ such that $\lim_{i \rightarrow \infty} \|f - g_i\| = \|f - M\|$, there exists $g \in M$ such that $\{g_i : i \in \mathcal{N}_+\}$ converges subsequentially to g (see, e.g., [23], p. 368). The following theorem is from [16].

Theorem 3.1 For all n, d positive integers, $\text{span}_n H_d$ is an approximatively compact subset of $(\mathcal{L}_p([0, 1]^d, \|\cdot\|_p))$ with $p \in [1, \infty)$.

The proof is based on an argument showing that any sequence of elements of $\text{span}_n H_d$ has a

subsequence that either converges to an element of $\text{span}_n H_d$ or to a Dirac delta distribution, and the latter case cannot occur when such a sequence "minimizes" a distance from some function in $\mathcal{L}_p([0, 1]^d)$.

It follows directly from the definitions that each approximatively compact subset is proximal.

Corollary 3.2 *For all n, d positive integers, $\text{span}_n H_d$ is a proximal subset of $(\mathcal{L}_p([0, 1]^d), \|\cdot\|_p)$ with $p \in [1, \infty)$.*

Thus, for any fixed number n , a function in $\mathcal{L}_p([0, 1]^d)$ has a best approximation among functions computable by a linear combination of n characteristic functions of half-spaces.

4 Uniqueness and continuity of a best approximation

Let M be a subset of a normed linear space $(X, \|\cdot\|)$ and let $\mathcal{P}(M)$ denote the set of all subsets of M . The set-valued mapping $P_M : X \rightarrow \mathcal{P}(M)$ defined by $P_M(f) = \{g \in M : \|f - g\| = \|f - M\|\}$ is called the *metric projection of X onto M* and $P_M(f)$ is called the *projection of f onto M* .

Let $F : X \rightarrow \mathcal{P}(M)$ be a set-valued mapping. A *selection* from F is a mapping $\phi : X \rightarrow M$ such that for all $f \in X$, $\phi(f) \in F(f)$. A mapping $\phi : X \rightarrow M$ is called a *best approximation operator* from X to M if it is a selection from P_M .

When M is proximal, then $P_M(f)$ is non-empty for all $f \in X$ and so there exists a best approximation mapping from X to M . The best approximation need not be unique. When it is unique, M is called a *Chebyshev set* (or "unicity" set). Thus M is Chebyshev if for all $f \in X$ the projection $P_M(f)$ is a singleton.

Recall that a normed linear space $(X, \|\cdot\|)$ is called *strictly convex* (also called "rotund") if for all $f \neq g$ in X with $\|f\| = \|g\| = 1$ we have $\|(f + g)/2\| < 1$. It is well known that for all $p \in (1, \infty)$, $(\mathcal{L}_p([0, 1]^d), \|\cdot\|_p)$ is strictly convex.

The following theorem from [13] implies for p in the open interval $(1, \infty)$ that if among best approximations to $\text{span}_n H_d$ (the existence of which is guaranteed by Corollary 3.2) there is a continuous one, then $\text{span}_n H_d$ must be a Chebyshev set.

Theorem 4.1 *In a strictly convex normed linear space, any subset with a continuous selection from its metric projection is Chebyshev.*

We shall combine this theorem with the following geometric characterization of Chebyshev sets with a continuous best approximation from [24].

Theorem 4.2 *In a Banach space with strictly convex dual, every Chebyshev subset with continuous metric projection is convex.*

It is well known that \mathcal{L}_p -spaces with $p \in (1, \infty)$ satisfy the assumptions of this theorem (since the dual of \mathcal{L}_p is \mathcal{L}_q where $1/p + 1/q = 1$ and $q \in (1, \infty)$) (see, e.g., [7], p. 160). Hence, to show the non-existence of a continuous selection, it is sufficient to verify that $\text{span}_n H_d$ is not convex.

Proposition 4.3 *For all n, d positive integers, $\text{span}_n H_d$ is not convex.*

Indeed, consider $2n$ parallel half-spaces with the characteristic functions $g_i(\mathbf{x}) = \vartheta(\mathbf{v} \cdot \mathbf{x} + b_i)$, where $0 > b_1 > \dots > b_{2n} > -1$ and $\mathbf{v} = (1, 0, \dots, 0) \in \mathcal{R}^d$. Then $\frac{1}{2} \sum_{i=1}^{2n} g_i$ is a convex combination of two elements of $\text{span}_n H_d$, $\sum_{i=1}^n g_i$ and $\sum_{i=n+1}^{2n} g_i$, but it is not in $\text{span}_n H_d$, since its restriction to the one-dimensional set $\{(t, 0, \dots, 0) \in \mathcal{R}^d : t \in [0, 1]\}$ has $2n$ discontinuities.

Summarizing results of this section and the previous one, we get the following corollary.

Corollary 4.4 In $(\mathcal{L}_p([0, 1]^d), \|\cdot\|_p)$ with $p \in (1, \infty)$ for all n, d positive integers there exists a best approximation mapping from $\mathcal{L}_p([0, 1]^d)$ to $\text{span}_n H_d$, but no such mapping is continuous.

Thus convenient properties of projection operators such as uniqueness and continuity are not satisfied by $\text{span}_n H_d$. These properties would allow one to estimate worst-case errors using methods of algebraic topology (see, e.g., [6]). In linear approximation theory, application of such methods shows that some sets of functions defined by smoothness conditions exhibit the curse of dimensionality: the approximants converge at rate $\mathcal{O}(1/\sqrt[n]{n})$, where d is the number of variables and n is the dimension of the approximating linear space (see, e.g., [20]). Our results show that these arguments are not applicable to approximation by $\text{span}_n H_d$.

5 Rates of approximation

Let $(X, \|\cdot\|)$ be a normed linear space and G be its subset, then G -variation (variation with respect to G) is defined as the Minkowski functional of the set $\text{cl conv}(G \cup -G)$, i.e.,

$$\|f\|_G = \inf\{c \in \mathcal{R}_+ : f/c \in \text{cl conv}(G \cup -G)\}.$$

Variation with respect to G is a norm on the subspace $\{f \in X : \|f\|_G < \infty\} \subseteq X$. The closure in its definition depends on the topology induced on X by the norm $\|\cdot\|$. When X is finite-dimensional, G -variation does not depend on the choice of a norm on X , since all norms on a finite-dimensional space are topologically equivalent.

Variation with respect to G has been introduced in [17] as an extension of the concept from [1] of H_d -variation called *variation with respect to half-spaces*. For functions of one variable, variation with respect to half-spaces coincides, up to a constant, with the notion of total variation studied in integration theory (see [1]). For G countable orthonormal, it coincides with l_1 -norm with respect to G (see [18]).

The following theorem from [17] is a reformulation of Maurey-Jones-Barron Theorem (see [2], [10], [21]) on estimates of rates of approximation of the order of $\mathcal{O}(1/\sqrt{n})$.

Theorem 5.1 Let $(X, \|\cdot\|)$ be a Hilbert space, G be its subset and $s_G = \sup_{g \in G} \|g\|$. Then for every $f \in X$ and for every positive integer n ,

$$\|f - \text{span}_n G\| \leq \sqrt{\frac{(s_G \|f\|_G)^2 - \|f\|^2}{n}}.$$

Corollary 5.2 For all positive integers d, n and for every $f \in (\mathcal{L}_2([0, 1]^d, \|\cdot\|_2))$,

$$\|f - \text{span}_n H_d\|_2 \leq \frac{\|f\|_{H_d}}{\sqrt{n}}.$$

Thus worst-case error in approximation of functions from the unit ball in H_d -variation by linear combinations of characteristic functions of n half-spaces of $[0, 1]^d$ is at most $1/\sqrt{n}$. Estimates derived from Theorem 5.1 are sometimes called "dimension-independent", which is misleading since with increasing number of variables, the condition of being in the unit ball in G -variation becomes more and more constraining. See [19] for examples of smooth functions with H_d -variation growing exponentially with the number of variables d . However, such exponentially growing lower bounds

on variation with respect to half-spaces are merely lower bounds on upper bounds on rates of approximation by $\text{span}_n H_d$, they do not prove that such functions cannot be approximated with faster rates than $\|f\|_{H_d}/\sqrt{n}$. Finding whether these exponentially large upper bounds are tight seems to be a difficult task related to some open problems in the theory of complexity of Boolean circuits.

Some insight into behavior of H_d -variation gives its geometric characterization derived in [19] using the Hahn-Banach Theorem.

Theorem 5.3 *Let $(X, \|\cdot\|)$ be a Hilbert space and G be its nonempty subset. Then for every $f \in X$, $\|f\|_G = \sup_{h \in S} \frac{|f \cdot h|}{\sup_{g \in G} |g \cdot h|}$, where $S = \{h \in X - G^\perp : \|h\| = 1\}$.*

Thus functions that are "almost orthogonal" to H_d (i.e., have small inner products with characteristic functions of half-spaces) have large H_d -variation.

6 Integral representation

The following theorem from [14] shows that a smooth real-valued function on \mathcal{R}^d with compact support can be represented as an integral combination of characteristic functions of half-spaces. By $H_{\mathbf{e},b}^-$ is denoted the half-space $\{\mathbf{x} \in \mathcal{R}^d : \mathbf{e} \cdot \mathbf{x} + b < 0\}$.

Theorem 6.1 *Let d be a positive integer and let $f : \mathcal{R}^d \rightarrow \mathcal{R}$ be compactly supported and $d+2$ -times continuously differentiable. Then*

$$f(\mathbf{x}) = \int_{S^{d-1} \times \mathcal{R}} w_f(\mathbf{e}, b) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) d\mathbf{e} db,$$

where for d odd

$$w_f(\mathbf{e}, b) = a_d \int_{H_{\mathbf{e},b}^-} \Delta^{k_d} f(\mathbf{y}) d\mathbf{y},$$

$k_d = (d+1)/2$, and a_d is a constant independent of f , while for d even,

$$w_f(\mathbf{e}, b) = a_d \int_{H_{\mathbf{e},b}^-} \Delta^{k_d} f(\mathbf{y}) \alpha(\mathbf{e} \cdot \mathbf{y} + b) d\mathbf{y},$$

where $\alpha(t) = -t \log |t| + t$ for $t \neq 0$ and $\alpha(0) = 0$, $k_d = (d+2)/2$, and a_d is a constant independent of f .

The assumption that f is compactly supported can be replaced by the weaker assumption that f vanishes sufficiently rapidly at infinity. The integral representation also applies to certain nonsmooth functions that generate tempered distributions.

By an approach reminiscent of Radon transform but based directly on distributional techniques from Courant and Hilbert [4], it was shown in [11] that if f is compactly supported function on \mathcal{R}^d with continuous d -th order partial derivatives, where d is odd, then f can be represented as

$$f(\mathbf{x}) = \int_{S^{d-1} \times \mathcal{R}} v_f(\mathbf{e}, b) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) d\mathbf{e} db,$$

where $v_f = a_d \int_{H_{\mathbf{e},b}} (D_{\mathbf{e}}^{(d)} f)(\mathbf{y}) d\mathbf{y}$, $a_d = (-1)^{k-1} (1/2) (2\pi)^{1-d}$ for $d = 2k+1$, $D_{\mathbf{e}}^{(d)} f$ is the directional derivative of f in the direction \mathbf{e} iterated d times, $d\mathbf{e}$ is the $(d-1)$ -dimensional volume element on S^{d-1} , and $d\mathbf{y}$ is likewise on a hyperplane. Although the coefficients v_f are obtained by integration over hyperplanes, while the w_f arise from integration over half-spaces, these coefficients can be shown to coincide by an application of the Divergence Theorem [3] p.423 to the half-spaces $H_{\mathbf{e},b}^-$. Theorem 6.1 extends the representation of [11] to even values for d and target functions f which are not compactly supported but which decrease sufficiently rapidly at infinity.

For $w \in \mathcal{L}_1(S^{d-1} \times \mathcal{R})$ and $f \in \mathcal{D}(\mathcal{R}^d)$ define

$$T_H(w)(\mathbf{x}) = \int_{S^{d-1} \times \mathcal{R}^d} w(\mathbf{e}, b) \vartheta(\mathbf{e} \cdot \mathbf{x} + b) d\mathbf{e} db,$$

$$S_H(f)(\mathbf{e}, b) = w_f(\mathbf{e}, b).$$

Theorem 6.1 shows that for each $f \in \mathcal{D}(\mathcal{R}^d)$, $T_H(S_H(f)) = f$. This theorem can be also used to estimate variation with respect to half-spaces by the \mathcal{L}_1 -norm of the weighting function $w_f = v_f$. It is shown in [11] that for any f to which the above representation applies,

$$\|f\|_{H_d} \leq \int_{S^{d-1} \times \mathcal{R}^d} |w_f(\mathbf{e}, b)| d\mathbf{e} db.$$

Combining this upper bound on H_d -variation with Corollary 5.2, we get a smoothness condition that defines sets of functions that can be approximated by $\text{span}_n H_d$ with rates of the order of $1/\sqrt{n}$.

Bibliography

1. Barron, A. R. (1992). Neural net approximation, in *Proceedings of the 7th Yale Workshop on Adaptive and Learning Systems* (pp. 69–72).
2. Barron, A. R. (1993). Universal approximation bounds for superposition of a sigmoidal function, *IEEE Transactions on Information Theory* **39**, 930–945.
3. Buck, R. C. (1965). *Advanced Calculus*, McGraw-Hill: New York.
4. Courant, R. and Hilbert, D. (1962). *Methods of Mathematical Physics*, vol. 2, Wiley: New York.
5. Cybenko, G. (1989). Approximation by superpositions of a single function, *Mathematics of Control, Signal and Systems* **2**, 303–314.
6. DeVore, R., Howard, R. and Micchelli, C. (1989). Optimal nonlinear approximation, *Manuscripta Mathematica* **63**, 469–478.
7. Friedman, A. (1982). *Foundations of Modern Analysis*, Dover: New York.
8. Gurvits, L. and Koiran, P. (1997). Approximation and learning of convex superpositions, *Journal of Computer and System Sciences* **55**, 161–170.
9. Hornik, K., Stinchcombe, M. and White, H. (1989). Multilayer feedforward networks are universal approximators, *Neural Networks* **2**, 251–257.
10. Jones, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Annals of Statistics* **20**, 608–613.

11. Kůrková, V., Kainen, P. C. and Kreinovich, V. (1997). Estimates of the number of hidden units and variation with respect to half-spaces, *Neural Networks* **10**, 1061–1068.
12. Kainen, P. C., Kůrková, V. and Vogt, A. (1999). Approximation by neural networks is not continuous, *Neurocomputing* **29**, 47–56.
13. Kainen, P. C., Kůrková, V. and Vogt, A. (2000). Geometry and topology of continuous best and near best approximations, *Journal of Approximation Theory* **105**, 252–262.
14. Kainen, P. C., Kůrková, V. and Vogt, A. (2000). An integral formula for Heaviside neural networks, *Neural Network World* **10** 313–319
15. Kainen, P. C., Kůrková, V. and Vogt, A. (2000). Best approximation by Heaviside perceptron networks. *Neural Networks* **13** 645–647.
16. Kainen, P. C., Kůrková, V. and Vogt, A. (2001). Best approximation by linear combinations of characteristic functions of half-spaces (submitted to J. of Approx. Theory).
17. Kůrková, V. (1997). Dimension-independent rates of approximation by neural networks, in *Computer-Intensive Methods in Control and Signal Processing: Curse of Dimensionality* (Eds. Warwick, K., Kárný, M.) (pp. 261–270). Birkhauser: Boston.
18. Kůrková, V., Sanguineti, M. (2001). Bounds on rates of variable-basis and neural network approximation, *IEEE Trans. on Information Theory* **47**, 2659–2665.
19. Kůrková, V., Savický, P. and Hlaváčková, K. (1998). Representations and rates of approximation of real-valued Boolean functions by neural networks, *Neural Networks* **11**, 651–659.
20. Pinkus, A. (1986). *n-Width in Approximation Theory*, Springer: Berlin.
21. Pisier, G. (1981). Remarques sur un resultat non publié de B. Maurey, in *Seminaire d'Analyse Fonctionnelle* I., n.12, Ecole Polytechnique, 1980–81.
22. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization of the brain, *Psychological Review* **65**, 386–408.
23. Singer, I. (1970). *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer: Berlin.
24. Vlasov, L. P. (1970). Almost convex and Chebyshev sets, *Math. Notes Acad. Sci. USSR* **8**, 776–779.
25. Zemanian, A. H. (1987). *Distribution Theory and Transform Analysis*, Dover: New York.